PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12N 15/10 — A1

(11) International Publication Number: WO 98/55657

(43) International Publication Date: 10 December 1998 (10.12.98)

(21) International Application Number: PCT/US98/11825

(22) International Filing Date: 5 June 1998 (05.06.98)

(30) Priority Data:
60/048,667 5 June 1997 (05.06.97) US

(71) Applicant *(for all designated States except US)*: CELLSTORE [US/US]; 26 Pigeon Hill Road, Weston, MA 02193 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: GIFFORD, David, K. [US/US]; 26 Pigeon Hill Road, Weston, MA 02193 (US).

(74) Agents: VINCENT, Matthew, P. et al.; Foley, Hoag & Eliot LLP, One Post Office Square, Boston, MA 02109 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS AND REAGENTS FOR INDEXING AND ENCODING NUCLEIC ACIDS

(57) Abstract

The current invention provides an accurate and reliable method for tagging gene sequences for future identification. The method uses a specific identification serial number that may be one or more characters, with each character being encoded by a distinct sequence of nucleic acids. These nucleic acids are referred to as the serial number nucleic acids. The distinct sequence of nucleic acids is attached to a given genetic sequence so that the genetic sequence will always be identifiable by one reading the serial number nucleic acids.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Methods and Reagents for Indexing and Encoding Nucleic Acids

Background of the Invention

Accurate and reliable identification of genetic sequences, including vectors and plasmids comprising such sequences, is. critical. As the field of biotechnology expands at an ever increasing pace, and the number of genetic sequences being discovered and used increases concurrently, it is imperative that those in the field be able to accurately store and identify sequences for further use. At the present time, biological depositories are available to store sequence specimens. When a deposit is made, a sample is assigned a deposit number which is then used in the future by one requesting a sample of the deposited sequence. One problem with such a system is that if a sample is misplaced or mislabeled, it is almost impossible to quickly and accurately determine what a given sample contains. An additional problem that arises with nucleic acid sequences is unauthorized transfer and the lack of means to track possession or ownership of a clone.

Prior to the current invention, the art was lacking a reliable and accurate way for tagging or identifying a given gene sequence. Therefore, the purpose of the present invention is to provide an identification serial number that can be assigned and attached to a given gene sequence and which will be available to identify the given gene sequence.

A still further aspect of the invention is to provide a vector comprising an identification serial number attached to a desired functional gene.

A further aspect of the present invention is to provide a secure system for tracking ownership or possession of a nucleic acid sequence comprising an identification serial number.

An additional aspect of the invention provides for a kit comprising an identification serial number available to one skilled in the art to join to a given nucleic acid sequence in order to tag a sequence for future identification.

Summary of the Invention

A method for the rapid identification of DNA clones is presented that identifies clones with unique serial numbers that can be easily determined. Vectors such as plasmids, YAC's and cosmids can all be numbered, and still remain fully compatible with conventional vector systems. Moreover, a cell having a desired functional gene or vector comprising a target sense nucleotide sequence, can be transfected with a vector comprising an identification serial number. Kits of manufactured vectors that contain specified, sequential, or random serial

numbers allow a user to mark a clone with a unique identification serial number when it is created. Once a clone is numbered it can be later positively identified using a simple serial number read-out method, preferably an array of labeled character detecting oligonucleotides. Such character detecting oligonucleotides can also be included in a kit.

Brief Description of the Drawings

Figure 1 depicts how a target sense nucleotide sequence (insert of interest) can be inserted into a serial numbered vector to create a vector having a serial number ("serial numbered vector") asset. The Read-Out portion of Figure 1 depicts how a serial numbered asset can have its identification serial number determined on a character detecting oligonucleotide read-out array.

Figure 2 depicts an example of an identification serial number that is engineered to encode a four character serial number (i.e., "5213"), where each character is one of the numerals 0 through 9 or the letters A through Z.

Figure 3 depicts how an identification serial number can be amplified using labeled primers, and how the resulting labeled identification serial number can be hybridized to an array of character detection oligonucleotides that are complementary to the nucleotide bases encoding each character to directly read-out the identification serial number.

Detailed Description of the Invention

The following terms are defined as follows: "Character" is used to refer to any number or letter or symbol used as, or part of, a serial number. Each character is encoded by a distinct sequence of nucleotide bases. "Character position" is used to refer to the position of each character in a given identification serial number. For example, the identification serial number "24B7" contains the character "4" in the second character position, and the character "B" in the third character position. "Identification serial number" is used to refer to a unique character, or unique set of characters, wherein each character is encoded by a distinct sequence of nucleotide bases. The terms "serial number nucleotide bases" and "SNNB sequence" are used interchangeably to refer to the nucleotide bases that encode the characters of a given serial number, to distinguish these nucleotide bases from the nucleotide bases that encode the sample genetic sequence being tagged. An "SNNB probe" is a nucleic acid, e.g., an oligonucleotide, which is complementary to at least a portion of an SNNB sequence, and which hybridizes to an SNNB sequence, typically to facilitate its detection. Generally, and SNNB probe will be the complement of a sequence for a given character position of the SNNB sequence. "Serial numbered asset" refers to the entire nucleic acid sequence including the serial number nucleotide bases and the nucleic acids that encode the

sample genetic sequence being tagged (e.g., a "functional gene sequence", or "sense sequence", or "target sense nucleotide sequence" or "insert of interest"). The term "nucleotide base" or "nucleotide bases" refers to both a single and/or a double stranded sequence of bases, and includes both DNA and RNA. "Character detection oligonucleotides" are oligonucleotides each of which comprises a sequence complementary to the sequence encoding a character. As used herein, the term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA), and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as equivalents, analogs of either RNA or DNA made from nucleotide analogs, and, as applicable to the embodiment being described, single (sense or antisense) and double-stranded polynucleotides. The term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked, and includes plasmids, cosmids or phages. Preferred vectors are those capable of autonomous replication. Also as used herein to describe nucleic acids, the terms "selectively hybridizes" and "specifically hybridizes" exclude the occasional randomly hybridizing nucleic acids.

In the practice of the present invention, a character is encoded by a unique set of nucleotide bases that differs from the sequence of any other character at at least one and preferably at multiple base positions. Preferably, the serial number nucleotide bases can be a non-native sequence of bases solely dedicated to identification. Alternatively, the serial number nucleotide bases can be incorporated into functional nucleotide sequences such as an antibiotic resistance gene by the careful choice of character encodings that do not destroy the function of the original native sequence. The use of functional bases for identification purposes can increase the effort required to remove a serial number. An identification serial number can optionally be flanked on one or both ends by fixed sequence(s) to enable the identification region to be amplified with PCR technology.

The identification serial number of a sample nucleotide clone, for example a DNA clone, can be determined through any of a number of sequencing techniques adaptable from the art. For instance, several methods for the semi- or fully automated sequencing of short nucleotide sequences have been developed, including minisequencing strategies (Pastinen et al., (1997) Genome Res. 7:606), multiplex reverse dot blots (Shuber et al., (1997) Hum. Mol. Genet 6:337), DNA chips (Fodor et al., (1991) Science 251:767), and the TaqMan approach (Livak et al., (1995) PCR Methods Appl 4:357).

For example, two methods for determining the sequence of the SNNB are by chemical cleavage, as disclosed by Maxim and Gilbert (1977), and by chain extension using ddNTPs, as disclosed by Sanger et al. (1977). In other embodiments, the sequence can be obtained by techniques utilizing capillary gel electrophoresis or mass spectroscopy. See, for example, U.S. Patent 5,003,059.

Alternatively, another method for determining the nucleotide sequence of an SNNB is to individually synthesize probes representing each possible sequence for each character position of an SNNB set. Thus, the entire set would comprise every possible sequence within the SNNB portion or some smaller portion of the set. By various deconvolution techniques, the identity of the probes which specifically anneal to the SNNB sequences can be determined.

An exemplary procedure would be to synthesize one or more sets of nucleic acid probes for detecting SNNB sequences simultaneously on a solid support. Preferred examples of a solid support include a plastic, a ceramic, a metal, a resin, a gel, and a membrane. A more preferred embodiment comprises a two-dimensional or three-dimensional matrix, such as a gel, with multiple probe binding sites, such as a hybridization chip as described by Pevzner et al. (J. Biomol. Struc. & Dyn. 9:399-410, 1991), and by Maskos and Southern (Nuc. Acids Res. 20:1679-84, 1992), both of which are herein specifically incorporated by reference.

Hybridization chips can be used to construct very large probe arrays which are subsequently hybridized with a target nucleic acid. Analysis of the hybridization pattern of the chip provides an immediate fingerprint identification of the SNNB sequence. Patterns can be manually or computer analyzed, but it is clear that positional sequencing by hybridization lends itself to computer analysis and automation. Algorithms and software have been developed for sequence reconstruction which are applicable to the methods described herein (Drmanac et al., (1992) Electrophoresis 13:566-73; P. A. Pevzner, J. Biomol. Struc. & Dyn. 7:63-73, 1989, both of which are herein specifically incorporated by reference).

For example, the identity of the SNNB sequence can be determined by annealing a solution of test sample nucleic acid including one or more SNNB sequences to a fixed array of character detection oligonucleotides (SNNB probes), where each column in the array preferably codes for one character of the identification serial number. Each fixed oligonucleotide has a nucleotide base sequence that is complementary to the nucleotide base sequence of a single character. Either the test sample nucleic acid or the fixed oligonucleotides can be labeled in such a fashion to permit read-out upon hybridization, e.g., by radioactive labeling or chemiluminescent labeling. Test nucleic acid can be labeled, for example, by using PCR to amplify the identification region of a DNA pool under test with PCR primers that are radioactive or chemiluminescent. Preferred detectable labels include a radioisotope, a stable isotope, an enzyme, a fluorescent chemical, a luminescent chemical, a chromatic chemical, a metal, an electric charge, or a spatial structure. There are many procedures whereby one of ordinary skill can incorporate detectable label into a nucleic acid. For example, enzymes used in molecular biology will incorporate radioisotope labeled substrate into nucleic acid. These include polymerases, kinases, and transferases. The labeling isotope is preferably, ^{32}P, ^{35}S, ^{14}C, or ^{125}I.

Moreover, recently Lockhart et al. (Nature Biotechnol. 14:1675,1996) published methods for the quantitative parallel measurement of cellular messenger RNA for gene sequences encoded on the chip solely from primary sequence data. RNAs present at a frequency of 1:300,000 were unambiguously detected with a quantitative assay spanning three to four orders of magnitude in concentration. Thus, a RNA sample including the SNNB sequence can be generated from the serial numbered asset by, for example, isolation of an mRNA which includes the SNNB sequence, or by use of T7 or T3 promoters flanking the SNNB sequence in a vector.

The labeled test nucleic acid is hybridized to the fixed array of oligonucleotides under conditions that are not permissive for test DNA/oligonucleotide duplexes when mismatches are present. The labeled nucleic acid may be directly or indirectly detected using scintillation fluid or a PhosphorImager, chromatic or fluorescent labeling, mass spectrometry or the like.

Other, more advanced methods of detection include evanescent wave detection of surface plasmon resonance of thin metal film labels such as gold, by, for example, the BIAcore sensor sold by Pharmacia, or other suitable biosensors. An exemplary plasmon resonance technique utilizes a glass slide having a first side on which is a thin metal film (known in the art as a sensor chip), a prism, a source of monochromatic and polarized light, a photodetector array, and an analyte channel that directs a medium suspected of containing an analyte, in this case an SNNB-containing nucleic acid, to the exposed surface of the metal film. A face of the prism is separated from the second side of the glass slide (the side opposite the metal film) by a thin film of refractive index matching fluid. Light from the light source is directed through the prism, the film of refractive index matching fluid, and the glass slide so as to strike the metal film at an angle at which total internal reflection of the light results, and an evanescent field is therefore caused to extend from the prism into the metal film. This evanescent field can couple to an electromagnetic surface wave (a surface plasmon) at the metal film, causing surface plasmon resonance. When an array of SNNB probes are attached to the sensor chip, the pattern of annealing to SNNB sequences produces a detectable pattern of surface plasmon resonance on the chip.

The pattern of annealing, e.g., of selective hybriziation, of the labeled test DNA to the oligonucleotide array or the test DNA to the labeled oligonucleotide array permits the identification serial number present in the original DNA clone to be directly read out. The detection array can include redundant oligonucleotides to provide integrated error checking. In general, the hybridization will be carried out under conditions wherein there is little background (non-specific) hybridization, e.g., the background level is at least one order of magnitude less than specific binding, and even more prefereably, at least two, three or four orders of magnitude less.

Additionally, the array can contain oligonucleotides that are known not to match any identification serial number as a negative control, and/or oligonucleotides that are known to match all identification serial numbers, e.g., primer flanking sequence, as a positive control.

In an certain embodiments, it is possible to include multiple identification serial numbers in a single asset (such as a cell line or virus) by choosing distinct flanking nucleotide sequences for each identification serial number that is introduced. To read out a single identification serial number, PCR primers that are complementary to its specific flanking sequences are used. The DNA that is amplified will be specific to the identification serial number selected.

In yet an alternative embodiment, DNA from multiple identification number loci can be hybridized against an anti-sense oligonucleotide array simultaneously. This can be accomplished if DNA from each loci has been prepared with a uniquely discernible tag. For example, each loci's oligonucleotide PCR primers can include unique moieties that result in loci specific color presentation during detection.

Oligonucleotides can be incorporated into a design array to determine interesting DNA family related motifs in a manner that is completely independent of serial number read-out. For example, the presence or absence of antibiotic resistance genes can be directly determined using oligonucleotides that are complementary to invariant portions of their coding region.

The nucleotide base sequences or oligonucleotide sequences of the present invention can be produced by conventional means known in the art, for example, recombinantly, chemically or mechanically (e.g., oligonucleotide synthesis machine). Various methods of chemically synthesizing polydeoxynucleotides are known, including solid-phase synthesis which has been fully automated in commercially available DNA synthesizers (See e.g., Itakura et al. U.S. Patent No. 4,598,049; Caruthers et al. U.S. Patent No. 4,458,066; and Itakura U.S. Patent Nos. 4,401,796 and 4,373,071, incorporated by reference herein).

In another aspect, the present invention provides a chip, such as a sensor chip, which provides an array of SNNB probes. Such arrays can be generated by various techniques known in the art. For instance, the arrays can be spatially synthesized utilizing light-directed chemical synthesis, such as photolithography or solid-phase synthesis. To illustrate the synthesis of one embodiment of the subject chips, synthetic linkers modified with photochemically removable protecting groups are attached to a glass substrate. Light is directed through a photolithographic mask to specific areas of the surface to produce localized photodeprotection. The first of a series of chemical building blocks --hydroxyl-protected deoxynucleosides, for example-- are incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. Next, light is directed to a different region of the substrate by a new mask, and the chemical cycle is

repeated. Highly efficient strategies can be used to synthesize any probe sequence at any discrete, specified location on the array in a minimum number of chemical steps. For example, the complete set of 4^n SNNB of length n, or any subset of this set, can be synthesized in only 4 x N chemical cycles. Thus, given a reference sequence, a DNA chip can be designed that consists of a highly dense array of complementary probes with no restriction on design parameters. The amount of nucleic acid information encoded on the chip in the form of different SNNB probes is limited only by the physical size of the array and the achievable lithographic resolution. Current bulk manufacturing methods allow for in excess of 409,000 polydeoxynucleotides to be synthesized on 1.28-cm by 1.28-cm chips.

Photolithography allows the construction of probe arrays with extremely high information content. Because the array is constructed on glass, it can be inverted and mounted in a temperature-controlled hybridization chamber. A sample SNNB sequence is fluorescently tagged and then injected into the chamber, where the target hybridizes to its complementary sequences on the array. Laser excitation enters through the back of the array, focused at the interface of the array surface and the target solution. Fluorescence emission is collected by a lens and passes through a series of optical filters to a sensitive detector. By simply scanning the laser beam or translating the array, or a combination of both, a quantitative two-dimensional fluorescence image of hybridization intensity is rapidly obtained. Commercial instrumentation for controlling the hybridization and scanning of the arrays, and software for image and data analysis have been developed. This approach requires only minute consumption of chemical reagents and minute preparations of biological samples.

Thus, in one embodiment, the subject system consists of chips arrayed with SNNB probes, a hybridization station to control hybridization with sample SNNB sequence, and a reader and software to access the chip data. At least two versions of commercial readers are available: a first-generation system from Molecular Dynamics as well as a recently released high-performance system from Hewlett-Packard. Moreover, chip production is now in a scaleable format. Affymax, for example, is now producing 5,000 to 10,000 chips per month.

In another embodiment, the identification of the serial number can be carried using molecular beacons (nucleic acid probes that only fluoresce when bound to their target sequence). See, for example, Piatek et al. (1998) Nature Biotechnology, 16:359-363; Tyagi et al. (1996) Nat. Biotechnol. 14:303; and Tyagi et al. (1998) Nat Biotechnol 16:49. To illustrate an embodiment of this technique, amplification of a sequence including the SNNB sequence is carried out in the presence of molecular beacons. Molecular beacons are typically hairpin-shaped, single-stranded oligonucleotides consisting of a probe sequence embedded within complementary sequences that form a hairpin stem. A fluorophore is covalently attached to one end of the oligonucleotide, and a nonfluorescent quencher is covalently attached to the other end. In the absence of a target, the fluorophore is held close to the

quencher and fluorescence cannot occur. When the probe binds to its target, the rigidity of the probe-target helix forces the stem to unwind, resulting in the separation of the fluorophore and quencher, and restoration of fluorescence. These probes can detect a number of different targets in the same solution (Tyagi et al. (1998) Nat Biotechnol 16:49). This is accomplished by constructing a different molecular beacon for each target and attaching a differently colored fluorophore to each. The probes are placed in the same amplification tube, and the color that develops indicates which targets were present. For example, two molecular beacons can be used, one specific for an SNNB sequence and labeled, e.g., with a green fluorophore and, the other specific for a control sequence and labeled, e.g., with a red fluorophore. The appearance of green fluorescence during amplification indicates the presence of the SNNB sequence, and red fluorescence indicates the presence of the control sequence. This approach can be used to analyze any DNA sequence of moderate length with single base pair accuracy. Piatek et al., supra.

The present invention also comprises vectors and host cells transformed to include the nucleotide base sequences of the invention. Suitable vectors, promoters, enhancers, and other expression control elements may be found in Sambrook et al. Molecular Cloning: A Laboratory Manual, second edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1989), incorporated by reference herein. Other suitable vectors, promoters, enhancers, and other expression and cellular elements are known to those skilled in the art. Moreover, methods of inserting a given nucleotide base sequence into a vector and methods of transfecting cells with said vector are known in the art. Several cellular systems are available to practice the present invention, for example yeast, bacterial and mammalian cell systems.

Host cells can be transformed to include the nucleotide base sequences of an identification serial number of the present invention using conventional techniques such as calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, or electroporation. Suitable methods for transfection and transformation may be found in Sambrook et al. *supra,* and other laboratory textbooks.

In another embodiment, viruses can be engineered to include identification serial numbers. In this embodiment, a virus' DNA or RNA genome includes an identification serial number. Techniques for engineering viral genome sequences are well known in the art and, as has already been described in the selection techniques for selecting nucleotide base sequences for characters, identification serial numbers in viruses can be placed either in biologically inactive or biologically active regions of the viral genome. The identification serial number sequence can be flanked at both ends by fixed nucleotide sequences as has been described to enable PCR amplification of the identification serial number loci.

A virus carrying an identification serial number can be introduced into a cell line using normal viral infection techniques that are well known in the art. Once a viral sequence has been integrated into the DNA of a host cell it can be read out using the techniques already presented for cells lines that contain identification serial numbers contained.

Identification serial numbers that are present in RNA form can be read out by first using reverse transcriptase to convert the RNA containing the identification serial number into DNA, and then PCR amplifying the resulting DNA using oligonucleotide primers that are complementary to the sequences that flank the identification serial number region of the virus. If reverse transcriptase is to be used to convert RNA to DNA for identification serial number readout, the RNA must be an appropriate substrate for reverse transcriptase and additional reverse transcriptase specific sequences may need to be added as is well known in the art outside of the fixed flanking regions.

The utilization of viruses that contain identification serial numbers have a number of additional applications. Vaccines based on viruses can have their vaccine type and date encoded into their serial number region, e.g. "Polio - 060497", and it would be possible to recover such serial identification serial numbers from individuals that had been immunized by the vaccine using PCR techniques as known in the art. This technique could also be used to monitor cell lines, animals, or individuals to see if they have been exposed to specifically labeled virus.

In another application or virus that carry serial numbers, viruses can be used to label existing cell lines with identification serial numbers using standard viral infection techniques.

In an embodiment of the present invention, an identification serial number of the present invention is provided comprised of at least one character. Each character is represented by a number, letter or symbol, and any combination of numbers, letters or symbols may be used in a given identification serial number. The identification serial number comprises a unique set of nucleotide bases which code for each character of the identification serial number. A unique set of nucleotide bases is provided to code for each character at each character site. For example, if an identification serial number is given as "53B5", the nucleotide bases coding for the character "5" at character position one are, in a preferred embodiment, unique and distinct from the nucleotide bases coding for the character "5" at character position four. Each unique nucleotide base sequence, encoding each character at each character site, will differ from other unique nucleotide bases sequences by at least one base position, most preferably at multiple base positions.

An identification serial number of the present invention may have any number of characters available at each character position, e.g., at least one, 5, 10, 15, 20, 25, 30, 35, 40, etc.. For example, each character position may have 36 separate nucleotide sequences which can be referred to by any of numbers 0 through 9, or any of letters A through Z. In such an

instance there would be 36 possible characters for each character position, and each of numbers 0 through 9 and letters, A through Z would be represented by a unique nucleotide base sequence depending upon the character position. For an identification serial number consisting of four characters, therefore, there would be 144 unique sequences (36 x 4) to represent each number or letter at each character site. Variations using the 144 unique sequences provides millions of different and unique identification serial number nucleotide base sequences.

Figure 2, for example, depicts construction of an identification serial number in a preferred embodiment. The nucleic acid sequence comprising the identification serial number consists of 80 bases arranged as four 20 base pair segments encoding four character positions, with two primer annealing sites of 20 base pairs each at each end. The primer annealing sites flank the character positions, which permits the identification site to be PCR amplified. Each character position is chosen from a unique set of 36 different 20 base pair sequences, each sequence differing from every other sequence in at least one base pair position. All of the sequences used to encode all of the characters differ from one another in this manner.

As shown in Figure 2, the fifth, second, first and third sequences from the respective character positions has been chosen to encode the serial number "5213". Note that the encoding of "5555" would include the fifth sequence from each character position, and would comprise four unique character sequences.

As discussed above, each character is encoded or represented by a unique nucleotide base sequence. The nucleotide base sequence can be any number of nucleotides in length, preferably I to 100, more preferably S to 50, and most preferably about 20 to 30 nucleotides in length. Each of the nucleotide base sequences, which individually represent each character, are joined together to form the identification serial number nucleotide sequence. The individual nucleotide base sequences may be joined in a manner known in the art, and preferably without any spacer sequence between the unique nucleotide sequences encoding each character. Additionally, the identification serial number nucleotide sequence may be flanked at either or both ends by fixed nucleotide sequences. Said fixed nucleotide sequence may be, for example, a sequence which will permit amplification (e.g., using PCR) of the identification serial number nucleotide sequence.

In addition to being unique, the nucleotide base sequences selected for each character, and the flanking sequences, are preferably chosen to be biologically inactive. Sequences that are biologically active (restriction site, promoters, MRNA polymerase start sites, etc.) are not ideally suitable for character encoding, although the use of such an active sequence may fall within the scope of the present invention. In addition, adjacent characters and the combination of all characters must be checked to make sure that biologically active sites do

not inadvertently arise from the combination of sequences. In yet another embodiment of the invention, for example, the identification serial number may be placed in a biologically active region of a plasmid or vector. In this embodiment, the nucleotide base sequences encoding each character must be chosen to not disturb the normal functioning of the active region that contains them. This can make the identification serial number nucleotide base sequence harder to remove and/or detect.

Other embodiments within the scope of the invention can use other identification serial number variations. For example, the number of bases used per symbol could be easily changed, as discussed above, from 20 to a different number; the four characters used in an identification serial number could be any variable number; and a different size or I (>cation of the primer annealing site could be employed.

The unique set of nucleotide bases encoding the characters of the identification serial number are provided as an oligonucleotide or in a vector system, as practiced and known by those skilled in the art. Upon receipt of the identification serial number, one may then, using methods known in the art, join the serial number nucleotide bases to a desired nucleotide sequence to be tagged (e.g., an active site or a target sense nucleotide sequence). Using methods known in the art, one may cleave a target sense nucleotide sequence (e.g., comprising a vector) at a position wherein the serial number nucleotide bases may be inserted (by, for example, sticky or blunt end techniques known in the art). Alternatively, one may cleave a vector comprising the serial number nucleotide bases at a position wherein the target sense nucleotide bases may be inserted. The identification serial number nucleotide sequence may, in one embodiment, be joined to the active or target sense nucleotide sequence so as to be in close proximity to each other, or more preferably adjacent to each other. Being in close proximity to each other diminishes the possibility that one may remove the identification serial number nucleotide base sequence without adversely affecting the active site.

Another alternative encompassed by the present invention is wherein one may maintain a cellular system (as described above) containing a vector comprising a target sense nucleotide sequence, and transfect the cell with a vector comprising the identification serial number nucleotide sequence. In this way the cellular system is maintained with both a vector comprising the target sense nucleotide sequence and a vector comprising the serial number nucleotide sequence.

One desiring to practice the invention may be provided with a kit comprising the identification serial number nucleotide base sequence, and may then tag an active sequence using techniques known in the art. Kits can also comprise character detection oligonucleotides. Alternatively, one may provide an active or target sequence to a depository, for example, where the active or target sequence may be tagged.

After a target sequence has been tagged, the characters of an identification serial number may be detected by permitting hybridization of the identification serial number nucleotide base sequences to one of a fixed array of character detection oligonucleotides. As discussed above, one skilled in the art would be able to create those conditions in which mismatches between the identification serial number nucleotide sequences and the fixed array of character detection oligonucleotides would not occur, and matches between the two would be detected by radioactivity or chemiluminescence. Once the labeled identification site is annealed or hybridized to the array of character detection oligonucleotides, the array is washed with a high stringency buffer, and unhybridized identification sites are eliminated. The remaining annealed/hybridized labeled identification sites permit direct read-out of the serial number. Visualization of the serial number on the array is accomplished using a label specific technique as is well known in the art.

For example, Figure 3 depicts how an identification serial number can be determined from a serial numbered asset. First, the identification site is PCR amplified with labeled primers. These primers can be radiolabeled or can be labeled with a nonradioactive moiety, such as a chemiluminescent moiety. The labeled identification site is then denatured and hybridized to one of an array of surface mounted character detection oligonucleotides. The array contains oligonucleotides with complementary sequences to all of the character encodings for all symbol positions. Thus, in our example shown in Figure 3, there would be 36 X 4 = 144 oligonucleotides in the fixed array of character detection oligonucleotides (which would permit 1,679,616 different serial numbers to be determined).

In another embodiment, the array includes multiple oligonucleotides with the same sequence that are spatially separated on the array for an internal control.

Oligonucleotides that are known not to match any identifier and all identifiers (e.g. an oligonucleotide complementary to the priming region) can be further included as negative and positive controls. The positive control indicates that an identified asset is being tested.

In the practice of the present invention a vector comprising the serial number nucleotide sequence may also comprise a selection gene sequence. A suitable selection gene sequence may, for example, be a drug resistance gene. This will enable one skilled in the art to maintain cells comprising the serial number nucleotide sequence and the drug resistance gene in a medium that will negatively select against those cells without the serial number nucleotide sequence and the drug resistance gene.

Figure 3 is an extension of the basic embodiment where other properties of a vector can be simultaneously determined. For example, if labeled primers are included in the PCR reaction that amplify other regions of interest, such as fragments of antibiotic markers, these markers can be detected at the same time that the sequence number is determined. Figure 3 depicts a labeled piece of the ampicillin gene sequence being present in the result of the PCR

reaction. This gene sequence is then directly detected by a complementary fixed oligonucleotide in the array.

Example

In a preferred embodiment of the invention, the Bluescript plasmid (Stratagene Corporation) is used as a vector system. The standard Bluescript plasmid is altered by including an identification serial number. The identification serial number included in the new vector will not disturb the existing biological activity of the plasmid, and furthermore does not introduce new activities. By not introducing the identification serial number in the middle of important coding regions, such as an antibiotic resistance gene or a multicloning site, the existing activities of the Bluescript plasmid are not disturbed. By choosing identification serial number nucleotide base sequences to be biologically inactive, as is described above, we ensure that the identification serial number nucleotide base sequences do not introduce any new biological activities.

In a preferred embodiment of the invention, identification serial numbered derivatives of Bluescript are made, each with a unique serial number. These manufactured vectors are made available to users of the vector system. As shown in Figure 1, when a user wishes to make a new clone that includes an identification serial number, the user selects a vector with a serial number that has not previously been used, and performs a routine cloning operation. The resulting identification serial numbered asset can be used and stored as would a normal Bluescript clone.

In an alternate embodiment, a clone that already includes a user vector can have an identification serial number added after the user cloning has taken place. In this embodiment, identification serial number nucleotide base sequences that encode serial numbers are manufactured with surrounding DNA that encodes an antibiotic resistance gene, and the entire construct has unique restriction sites on either end. This identification construct can be cloned into an existing user clone. Antibiotic selection can be used to select user clones that incorporate the identification serial number. Alternately, an antibiotic marker does not have to be used, and user clones that have taken up the identification serial number can be identified by reading out their serial numbers as described above.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, numerous equivalents to the specific procedures described herein. Such equivalents are considered to be within the scope of this invention and are covered by the following claims.

Claims

1 An identification serial number comprised of at least one character, wherein each of said at least one character is encoded by a sequence of nucleotide bases.

2. The identification serial number of claim 1 wherein said sequence of nucleotide bases comprise DNA.

3. The identification serial number of claim I wherein each character represents a unique sequence of bases wherein each of said sequences of bases which encode each character differs from each other by at least one base position.

4. The identification serial number of claim 3 wherein each of said sequences of bases differs from each other at multiple base positions.

5. The identification serial number of claim I wherein each character of the serial number can represent one of a fixed number of unique sequences of bases wherein each of said sequences of bases differs from each other by at least one base position.

6. The identification serial number of claim 5 wherein each of said sequences of bases differs from each other at multiple base positions.

7. The identification serial number of claim 3 wherein each character of the serial number can represent one of a fixed number of unique sequences of bases wherein each of said sequences of bases differs from each other by at least one base position.

8. The identification serial number of claim 7 wherein each of said sequences of bases differs from each other at multiple base positions.

9. The identification serial number of claim I wherein each character of the serial number is represented by one of 36 sequences of bases.

10. The identification serial number of claim 5 wherein said fixed number of sequences is 36.

11. The identification serial number of claim I which is biologically inactive.

12. A vector comprising an identification serial number of claim 1.

13. A vector comprising an identification serial number of claim 3.

14. The vector of claim 12 further comprising at least one fixed nucleotide sequence which enables amplification of said identification serial number.

15. The vector of claim 14 wherein the identification serial number is flanked on both ends by at least one fixed nucleotide sequence.

16. The vector of claim 13 further comprising at least one fixed nucleotide sequence which enables amplification of said identification serial number.

17. The vector of claim 16 wherein -:he identification serial number is flanked on both ends by at least one fixed nucleotide sequence.

18. The vector of claim 14 further comprising an active site.

19. The vector of claim 18 wherein the identification serial number is in close proximity to the active site.

20. The vector of claim 18 wherein the identification serial number is adjacent to the active site.

21. A kit comprising at least one vector of claim 12 wherein each of said at least one vector comprises a unique identification serial number.

22. The kit of claim 21 further comprising character detection oligonucleotides.

23. The kit of claim 22 further comprising negative and positive control oligonucleotides.

24. A serial numbered asset comprising the vector of claim 12 and a sense sequence which comprises a gene of interest.

25. A serial numbered asset comprising the vector of claim 13 and a sense sequence which comprises a gene of interest.

26. A method of tagging a sense sequence comprising providing a target sense nucleotide sequence to be tagged; providing an identification serial number of claim 1; and joining said identification serial number to said sense sequence.

27. A method of tagging a sense sequence comprising providing a target sense nucleotide sequence to be tagged; providing an identification serial number of claim 3; and joining said identification serial number to said sense sequence.

28. A method of tagging a sense sequence comprising providing a target sense nucleotide sequence to be tagged; providing a vector of claim 12; and incorporating said sense sequence into said vector.

29. A method of tagging a sense sequence comprising providing a target sense nucleotide sequence to be tagged; providing a vector of claim 13; and incorporating said sense sequence into said vector.

30. A method of tagging a cell comprising a vector comprising a target sense nucleotide sequence, comprising providing a vector of claim 12; and transfecting said vector of claim 12 into said cell.

31. A method of tagging a cell comprising a vector comprising a target sense nucleotide sequence, comprising providing a vector of claim 13; and transfecting said vector of claim 13 into said cell.

32. The vector of claim 12 further comprising a selection gene sequence.

33. The vector of claim 32 wherein said selection gene sequence encodes for drug resistance.

34. The vector of claim 13 further comprising a selection gene sequence.

35. The vector of claim 34 wherein said selection gene sequence encodes for drug resistance.

36. A method of detecting the characters of an identification serial number of claim 1, said method comprising hybridizing said identification serial number to a fixed array of character detection oligonucleotides.

37. The method of claim 36 wherein said identification serial number is labeled for detection.

38. The method of claim 36 wherein said character detection oligonucleotides are labeled for detection.

39. A method of detecting the characters of an identification serial number contained in a vector of claim 12, said method comprising annealing said identification serial number to a fixed array of character detection oligonucleotides.

40. The method of claim 39 wherein said identification serial number is labeled for detection.

41. The method of claim 39 wherein said character detection oligonucleotides are labeled for detection.

Figure 1

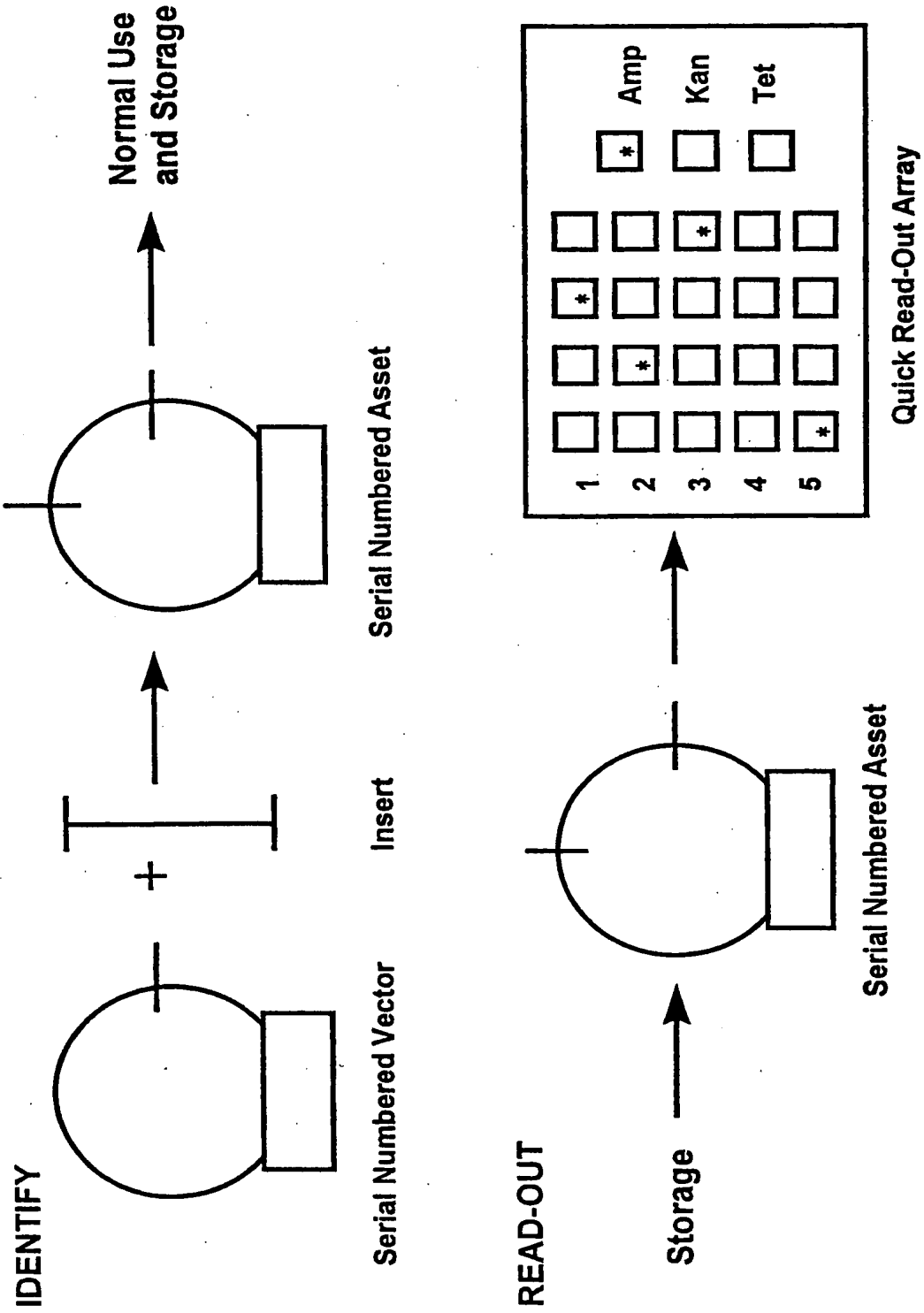

IDENTIFY
Serial Numbered Vector
+
Insert
Serial Numbered Asset
Normal Use and Storage
READ-OUT
Storage
Serial Numbered Asset
Amp
Kan
Tet
1
2
3
4
5
Quick Read-Out Array

Figure 2

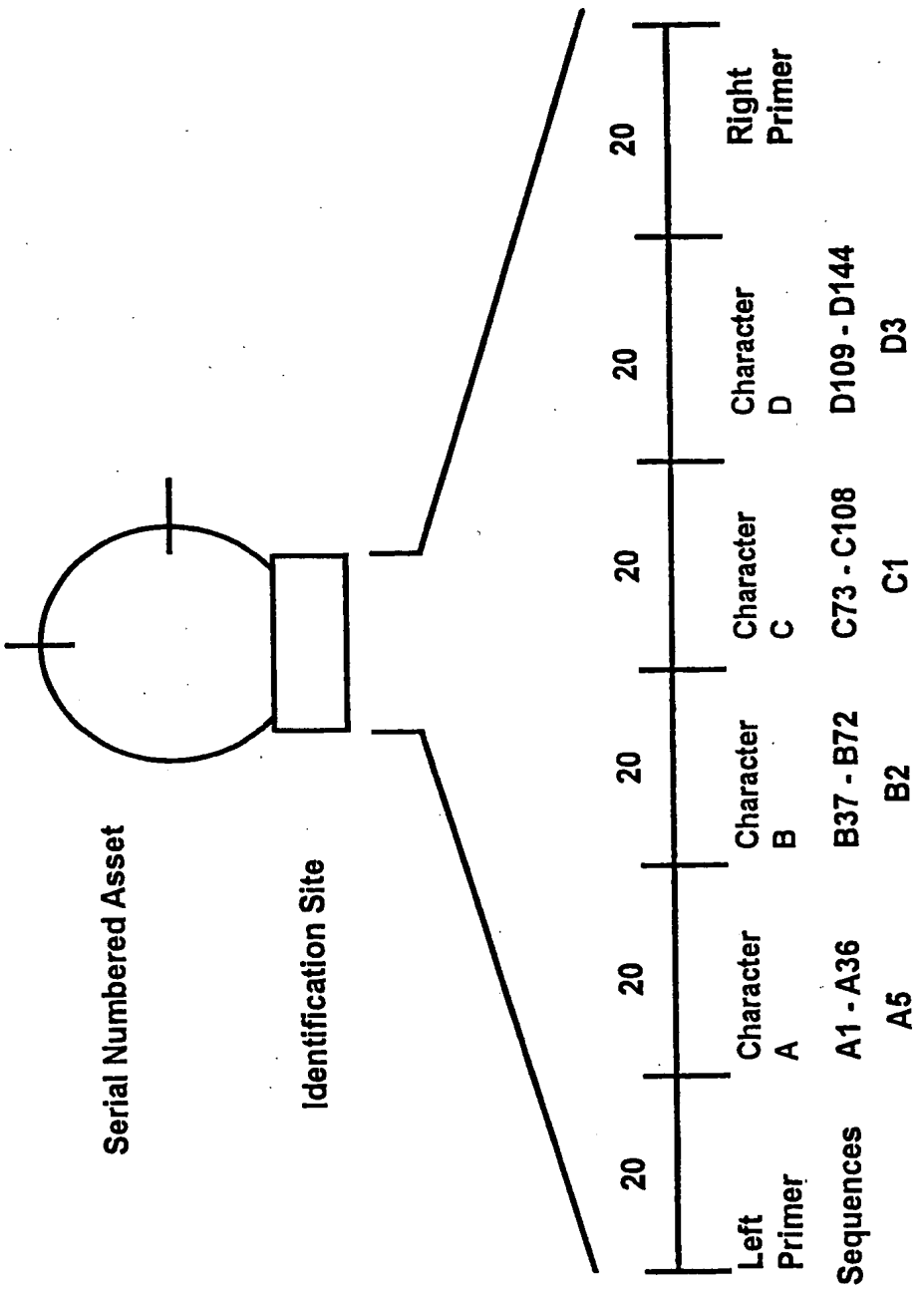
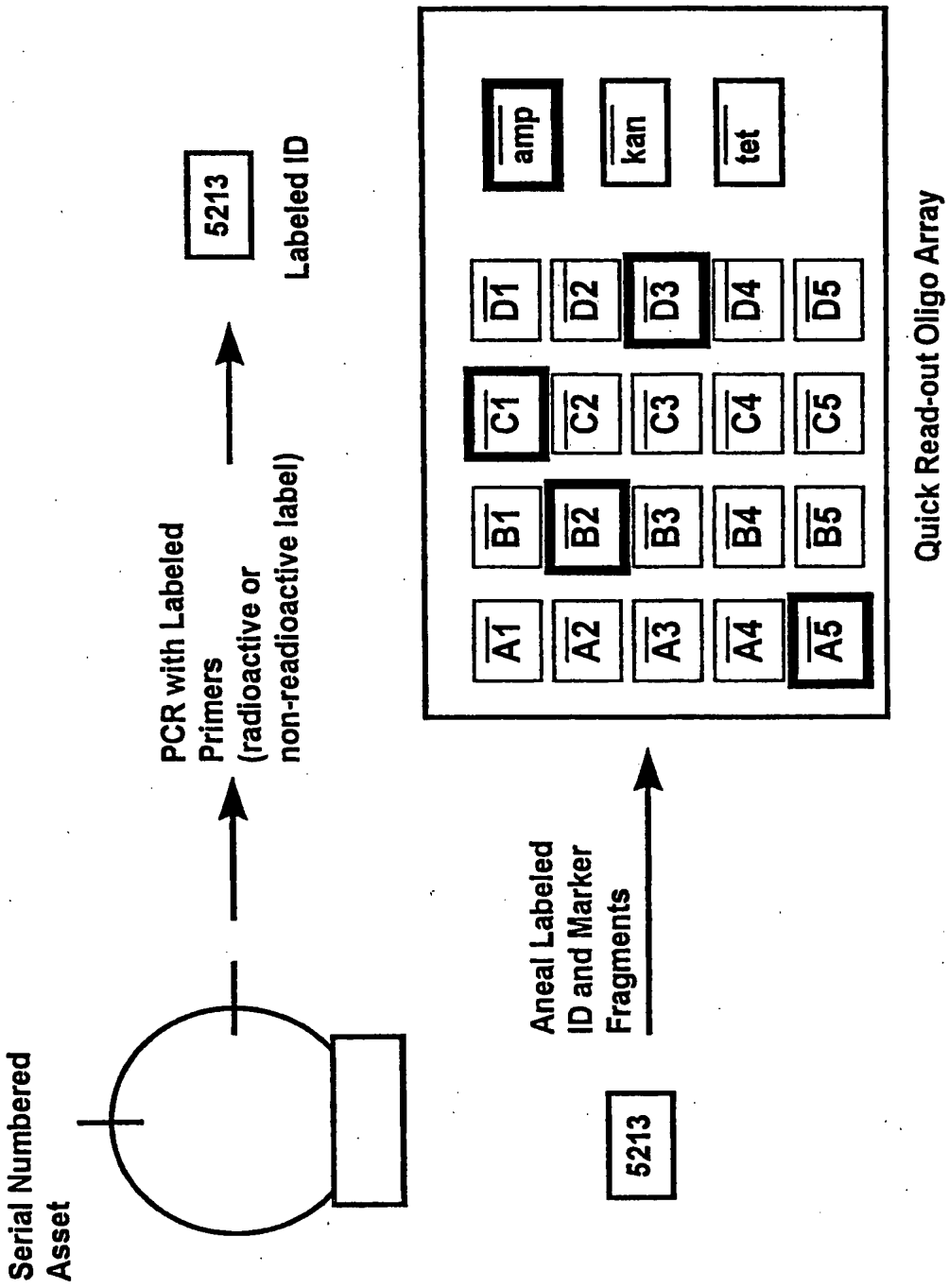

Serial Numbered Asset
Identification Site
Left Primer
20
Character A
20
Character B
20
Character C
20
Character D
20
Right Primer
Sequences
A1 - A36
A5
B37 - B72
B2
C73 - C108
C1
D109 - D144
D3

Figure 3


Serial Numbered Asset
PCR with Labeled Primers (radioactive or non-readioactive label)
5213
Labeled ID
Aneal Labeled ID and Marker Fragments
5213
amp
kan
tet
A1
A2
A3
A4
A5
B1
B2
B3
B4
B5
C1
C2
C3
C4
C5
D1
D2
D3
D4
D5
Quick Read-out Oligo Array

INTERNATIONAL SEARCH REPORT

Inte ional Application No
PCT/US 98/11825

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68 C12N15/10

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	WO 96 17954 A (PABIO ;ALESTROEM PETER (NO)) 13 June 1996 see the whole document ---	1-29, 36-41 31-35
X Y	WO 96 12014 A (LYNX THERAPEUTICS INC) 25 April 1996 see the whole document ---	1-11,30 31-35
X	US 5 149 625 A (CHURCH GEORGE M ET AL) 22 September 1992 see the whole document --- -/--	1,2,11, 12,14, 15, 18-26,28

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publicationdate of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of theinternational search	Date of mailing of the international search report
15 October 1998	02/11/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl. Fax: (+31-70) 340-3016	Hagenmaier, S

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inte ional Application No
PCT/US 98/11825

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN AND KRAMER: "OLIGONUCLEOTIDE ARRAYS:NEW CONCEPTS AND POSSIBILITIES" BIOTECHNOLOGY, vol. 12, 1994, pages 1093-1099, XP002080897 see the whole document	36-41
P,X	EP 0 799 897 A (AFFYMETRIX INC) 8 October 1997 see the whole document	1-41

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

information on patent family members

Int. ional Application No
PCT/US 98/11825

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9617954 A	13-06-1996	AU 3992395 A	26-06-1996
		CA 2206486 A	13-06-1996
		EP 0795029 A	17-09-1997
		NO 972610 A	07-08-1997
WO 9612014 A	25-04-1996	US 5604097 A	18-02-1997
		AU 3946195 A	06-05-1996
		AU 4277896 A	06-05-1996
		CA 2202167 A	25-04-1996
		CZ 9700866 A	17-09-1997
		EP 0786014 A	30-07-1997
		EP 0793718 A	10-09-1997
		FI 971473 A	04-06-1997
		JP 10507357 T	21-07-1998
		NO 971644 A	02-06-1997
		WO 9612039 A	25-04-1996
		US 5695934 A	09-12-1997
		US 5635400 A	03-06-1997
		US 5654413 A	05-08-1997
US 5149625 A	22-09-1992	US 4942124 A	17-07-1990
		CA 1339727 A	17-03-1998
		EP 0303459 A	15-02-1989
		JP 1137982 A	30-05-1989
		JP 2665775 B	22-10-1997
EP 0799897 A	08-10-1997	NONE	